

What is claimed is:

1. A computer program product for efficiently extracting data from a data stream, the computer program product embodied on one or more computer-readable media and comprising:
computer-readable program code means for defining one or more data extraction rules, each of the rules comprising one or more rule components;

computer-readable program code means for defining one or more output document templates for storing extracted data, wherein each of the templates comprises one or more tags which are hierarchically structured and wherein each template is to be associated with one or more of the data extraction rules;

computer-readable program code means for associating at least one of the templates with at least one of the rules;

computer-readable program code means for storing the rules, the templates, and the associations;

computer-readable program code means for monitoring at least one data stream for arrival of incoming data;

computer-readable program code means for comparing the incoming data to selected ones of the stored rules until detecting a matching rule;

computer-readable program code means for extracting data from the incoming data, upon detecting the matching rule, according to the matching rule; and

computer-readable program code means for storing the extracted data in an extensible document which is created according to the tags and structure of a selected one of the templates that is associated with the matching rule.

1 2. The computer program product according to Claim 1, wherein the computer-readable
2 program code means for associating further comprises computer-readable program code means
3 for associating the rule components of a particular rule with the tags of a particular template.

1 3. The computer program product according to Claim 1, further comprising computer-
2 readable program code means for transforming the extracted data in the extensible document into
3 another notation.

1 4. The computer program product according to Claim 1, further comprising computer-
2 readable program code means for transforming the extracted data in the extensible document into
3 another format.

1 5. The computer program product according to Claim 1, wherein the extensible document is
2 an Extensible Markup Language ("XML") document.

1 6. The computer program product according to Claim 1, wherein the components of selected
2 ones of the rules specify textual patterns.

1 7. The computer program product according to Claim 1, wherein the components of selected
2 ones of the rules specify data element and attribute patterns.

1 8. The computer program product according to Claim 1, wherein the components of selected
2 ones of the rules specify a combination of textual patterns and data element and attribute patterns.

1 9. A system for efficiently extracting data from a data stream, comprising:
2 means for defining one or more data extraction rules, each of the rules comprising one or
3 more rule components;
4 means for defining one or more output document templates for storing extracted data,
5 wherein each of the templates comprises one or more tags which are hierarchically structured and
6 wherein each template is to be associated with one or more of the data extraction rules;
7 means for associating at least one of the templates with at least one of the rules;
8 means for storing the rules, the templates, and the associations;
9 means for monitoring at least one data stream for arrival of incoming data;
10 means for comparing the incoming data to selected ones of the stored rules until detecting
11 a matching rule;
12 means for extracting data from the incoming data, upon detecting the matching rule,
13 according to the matching rule; and
14 means for storing the extracted data in an extensible document which is created according
15 to the tags and structure of a selected one of the templates that is associated with the matching
16 rule.

1 10. The system according to Claim 9, wherein the means for associating further comprises
2 means for associating the rule components of a particular rule with the tags of a particular

3 template.

1 11. The system according to Claim 9, further comprising means for transforming the extracted
2 data in the extensible document into another notation.

1 12. The system according to Claim 9, further comprising means for transforming the extracted
2 data in the extensible document into another format.

1 13. The system according to Claim 9, wherein the extensible document is an Extensible
2 Markup Language (“XML”) document.

1 14. The system according to Claim 9, wherein the components of selected ones of the rules
2 specify textual patterns.

1 15. The system according to Claim 9, wherein the components of selected ones of the rules
2 specify data element and attribute patterns.

1 16. The system according to Claim 9, wherein the components of selected ones of the rules
2 specify a combination of textual patterns and data element and attribute patterns.

1 17. A method for efficiently extracting data from a data stream, comprising the steps of:
2 defining one or more data extraction rules, each of the rules comprising one or more rule

3 components;

4 defining one or more output document templates for storing extracted data, wherein each
5 of the templates comprises one or more tags which are hierarchically structured and wherein each
6 template is to be associated with one or more of the data extraction rules;

7 associating at least one of the templates with at least one of the rules;

8 storing the rules, the templates, and the associations;

9 monitoring at least one data stream for arrival of incoming data;

10 comparing the incoming data to selected ones of the stored rules until detecting a
11 matching rule;

12 extracting data from the incoming data, upon detecting the matching rule, according to the
13 matching rule; and

14 storing the extracted data in an extensible document which is created according to the tags
15 and structure of a selected one of the templates that is associated with the matching rule.

1 18. The method according to Claim 17, wherein the associating step further comprises the
2 step of associating the rule components of a particular rule with the tags of a particular template.

1 19. The method according to Claim 17, further comprising the step of transforming the
2 extracted data in the extensible document into another notation.

1 20. The method according to Claim 17, further comprising the step of transforming the
2 extracted data in the extensible document into another format.

1 21. The method according to Claim 17, wherein the extensible document is an Extensible
2 Markup Language (“XML”) document.

1 22. The method according to Claim 17, wherein the components of selected ones of the rules
2 specify textual patterns.

1 23. The method according to Claim 17, wherein the components of selected ones of the rules
2 specify data element and attribute patterns.

1 24. The method according to Claim 17, wherein the components of selected ones of the rules
2 specify a combination of textual patterns and data element and attribute patterns.

1 25. The method according to Claim 17, wherein the data stream is a legacy host stream
2 containing one or more presentation spaces.

1 26. The method according to Claim 17, wherein the data stream is sent between peer
2 applications.

1 27. The method according to Claim 26, wherein the data stream contains one or more
2 Extensible Markup Language (“XML”) documents.

1 28. The method according to Claim 17, wherein the data stream contains one or more Web
2 pages.